BOIES SCHILLER FLEXNER LLP
David Boies (*pro hac vice*)
333 Main Street
Armonk, NY 10504
(914) 749-8200
dboies@bsfllp.com

LIEFF CABRASER HEIMANN & BERNSTEIN, LLP
Rachel Geman (*pro hac vice*)
250 Hudson Street, 8th Floor
New York, New York 10013-1413
(212) 355-9500
rgeman@lchb.com

CAFFERTY CLOBES MERIWETHER & SPRENGEL LLP
Bryan L. Clobes (*pro hac vice*)
135 S. LaSalle Street, Suite 3210
Chicago, IL 60603
(312) 782-4880
bclobes@caffertyclobes.com

Counsel for Individual and Representative Plaintiffs and the Proposed Class

(additional counsel included below)

JOSEPH SAVERI LAW FIRM, LLP
Joseph R. Saveri (SBN 130064)
601 California Street, Suite 1505
San Francisco, California 94108
(415) 500-6800
jsaveri@saverilawfirm.com

DICELLO LEVITT LLP
Amy Keller (*pro hac vice*)
10 North Dearborn Street, Sixth Floor
Chicago, Illinois 60602
(312) 214-7900
akeller@dicellolevitt.com

Matthew Butterick (SBN 250953)
1920 Hillhurst Avenue, #406
Los Angeles, CA 90027
(323) 968-2632
mb@buttericklaw.com

UNITED STATES DISTRICT COURT

NORTHERN DISTRICT OF CALIFORNIA

SAN FRANCISCO DIVISION

RICHARD KADREY, et al.,

Individual and Representative Plaintiffs,

v.

META PLATFORMS, INC.,

Defendant.

Case No. 3:23-cv-03417-VC

PLAINTIFFS' REPLY TO MOTION FOR PARTIAL SUMMARY JUDGMENT AND OPPOSITION TO META'S MOTION FOR PARTIAL SUMMARY JUDGMENT

TABLE OF CONTENTS

TABLE OF AUTHORITIES

Page(s)

Cases

Statutes

Rules

Other Authorities

Meta frames this case as presenting "a question of existential importance to the future of generative artificial intelligence ('AI') development in the United States." Meta Br. at 1. But that proclamation cannot, as a matter of law, relieve Meta of liability for its unprecedented online piracy of copyrighted literary works or excuse Meta's other infringements under the guise of fair use.

It is now undisputed that Meta torrented tens of millions of pirated books and other copyrighted works, including over 650 copies of Plaintiffs' Books, for free and without consent from the rightsholders because it did not want to pay for them. Over a century of precedent holds that pirating intellectual property creates liability for infringement. When a person does this, they can be held civilly and criminally liable. But Meta asks this Court to hold that its conduct is somehow different—that unlawful conduct by one of the world's largest tech companies is somehow legitimized by doing it on a massive scale because of "existential" technological needs. According to Meta, it *must* be allowed to infringe copyrighted works *en masse* because there's no other way to train LLMs. Yet even Meta knows pirating books is not necessary to develop LLMs; Meta itself developed "non-LibGen" models trained on other material. Ex. 61.[1] Meta's mass copyright infringement therefore is not inherent to LLM development—it just served to make Meta's models marginally better relative to some of its competitors. Far from being a referendum on emerging technology, this case is simply about whether Meta must comply with existing law in developing its commercial products, or whether Meta may indiscriminately use, for its own financial gain and without compensation, all intellectual property protected by U.S. copyright law.

Meta tries to avoid grappling with the facts here by drawing a false equivalency between its conduct and the conduct at issue in *Authors Guild v. Google, Inc.*, 804 F.3d 202 (2d Cir. 2015) ("*Google Books*"). That case, however, which itself "test[ed] the boundaries of fair use," *id.* at 206, is a far cry from Meta's reality here. *Google Books* was the result of a legitimate collaborative partnership between Google and "a number of the world's major research libraries." *Id.* at 208. In

[1] Plaintiffs commence new exhibits at number 97; all citations to exhibits 1-96 refer to the materials accompanying Plaintiffs' affirmative motion for partial summary judgment, ECF Nos. 472-75. Unless indicated otherwise, all exhibit citations are to the declarations of Maxwell V. Pritt.

the pursuit of the expansion of collective knowledge, Google executed "bi-lateral agreements" with those libraries to "create[] an index of the machine-readable text" of books, "the vast majority of [which] are non-fiction, and most [] out of print." *Id.* at 208. And when users search the Google Books index, they are directed to links to buy the queried books whose snippets are displayed. Google also receives no payment for this function, whether through advertising or "finders' fees" from eventual book purchases. *Id.* at 209. The Google Books project thus resulted in no meaningful market substitution for the original copyrighted works. *Id.* at 224. And the *Google Books* court even lauded Google's "impressive security measures" *against* the risk of data piracy. *Id.* at 228.

Little of what Meta did resembles *Google Books*. Instead of buying or licensing copyrighted works, Meta torrented tens of millions of them from pirated databases so notorious that they're frequently targeted by the FBI and enjoined by federal courts. Instead of directing users *to* the copied works, Meta *cannibalized* licensing revenue from those works by depriving authors of compensation in the exploding market for AI training data. And far from targeting purely factual works, Meta indiscriminately ingested the entire copyrighted universe, fiction and non-fiction alike, ironically lamenting that it had "only collected 5% of the world book content" from Anna's Archive, and without paying a single cent. Ex. 97, Meta_Kadrey_00211192, at -195.

No court has come close to holding that such egregious conduct could constitute "fair use." This Court should not be the first. Accordingly, Plaintiffs respectfully request the Court grant Plaintiffs' partial motion for summary judgment and deny Meta's motion for summary judgment.

I. AS A MATTER OF LAW, META'S PIRACY OF COPYRIGHTED WORKS IS NOT FAIR USE

The original method of reproduction of copyrighted material (*i.e.*, how the infringer made unauthorized copies) bears directly on fair use, and no method has been as resoundingly rejected as pirating content through peer-to-peer ("P2P") networks. Take Meta's word for it. Even after Plaintiffs filed this case in 2023, Meta submitted comments to the U.S. Copyright Office stating:

> [O]ne of the most critical elements of the balance between rightsholder interests and innovation is the doctrine of fair use. Courts have applied that doctrine to lay the legal groundwork for revolutionary new technologies like internet search, ***while declining to sanction more exploitative technologies, like unauthorized file sharing*** and unlicensed media clip services.

Comments of Meta Platforms, U.S. Copyright Office Dkt. No. 2023-06, at 11 (filed Oct. 30, 2023) (emphasis added).[2]

There is no dispute that Meta downloaded hundreds of terabytes ("TB") of copyrighted works from known pirated databases, including several hundred copies of Plaintiffs' Books, and that Meta torrented the lion's share of that content. *See* Ex. 72 at 47, Table 2. And there now is no dispute that Meta also *distributed* a massive share of that data, contributing invaluable bandwidth, storage, and processing power to Meta's fellow digital pirates—Meta's own expert found that Meta's Amazon Web Services ("AWS") cost and usage data shows that "peers could have received from Meta [] ***approximately 30%*** [*i.e.*, ***over 40 TB***] of the data that Meta downloaded." Dkt. 492 at ¶ 29; Ex. 163, Frederiksen-Cross Second Rebuttal Report, at 28-29 n. 56 (emphasis added).[3] Meta ignores these points and at best highlights a dispute over a single immaterial fact: whether *Plaintiffs'* Books were among the 40 TB of copyrighted content that Meta admits it distributed online in a three-month period in 2024. Meta Br. at 34-35. That evidence is equal parts irrelevant and available only by inference. Moreover, Meta "destroyed" critical parts of its torrenting logs that would provide additional circumstantial evidence. *See* Section I.B.1, *infra*. Because no reasonable jury could find Meta's mass piracy was fair use, Plaintiffs' Motion should be granted.

A. Meta's Attempt To Distinguish the Uniform Body of Precedent Rejecting Fair Use for Illegally Acquired Copyrighted Material Is Unavailing.

1. Unmitigated piracy continues to be a categorically unfair use.

In their Motion, Plaintiffs noted they could not locate a single case that sanctioned as fair use unmitigated piracy such as the use of P2P file sharing networks to copy (or distribute) works without permission. Pltfs' Br. at 23-27. In response, Meta cites ***no such cases***, because none exist.

---

[2] Available at https://www.regulations.gov/comment/COLC-2023-0006-9027.

[3] Frederiksen-Cross's second rebuttal report also corrects duplication in her prior report's 267.4 TB calculation of what Meta downloaded via torrent from LibGen, Z-Lib, and IA between April and July 2024. The new report calculates downloading of 134.6 TB based on Meta's AWS billing data between April and July 2024 while also noting Meta uploaded 40.42 TB during that same time period. The amount of copyrighted data that Meta torrented from just these three pirated databases is astonishing, and these figures do not even include the additional torrenting Meta did from LibGen in 2022 and 2023 and Meta's torrenting from other pirated databases.

Meta and its amici attempt to distinguish this case from the many P2P file sharing cases that uniformly deny fair use to defendants engaged in this form of infringement. According to Meta, its "use was fair irrespective of its method of acquisition." Meta Br. at 35. But Meta does not cite a single case—either within or outside the P2P context—where a defendant used a copy of a work it stole and successfully mounted a fair use defense. Nor has Meta identified any cases where this form of unmitigated digital piracy has received anything but back-of-the-hand treatment from the courts. To the contrary, "[i]n some cases, no analysis is required; it is obvious, for example, that downloading and distributing copyrighted music via peer-to-peer systems does not constitute fair use." *In re DMCA § 512(H) Subpoena to Twitter, Inc.*, 608 F. Supp. 3d 868, 879 (N.D. Cal. 2022) (Chhabria, J.); *United States v. Slater*, 348 F.3d 666, 669 (7th Cir. 2003) ("It is preposterous to think that Internet piracy is authorized by virtue of the fair use doctrine"); *see also Glacier Films (USA), Inc. v. Turchin*, 896 F.3d 1033, 1043 (9th Cir. 2018) (calling fair use a "baseless affirmative defense[]" where defendant "conced[ed] that he downloaded [a movie]" via BitTorrent). With it entirely undisputed that Meta torrented well over 134.6 TB of copyrighted text from known pirated websites, the Court can simply follow all analogous precedent and reject fair use outright.

2. Meta's method of reproduction nevertheless bears heavily on fair use.

Plaintiffs' Motion stands for a simple, uncontroversial proposition: fair use cannot apply if the copyrighted work was illegally obtained in the first instance. Pltfs' Br. at 22-23. But even if unmitigated piracy is subject to the fair-use factors, the method of acquisition still bears heavily on the analysis: Where a defendant acquires a work by lying or stealing, it is fundamentally *unfair*. *See Nunez v. Caribbean Int'l News Corp.*, 235 F.3d 18, 23 (1st Cir. 2000) ("unlawful acquisition of the copyrighted work generally weighs against a finding of fair use"). Here, Meta relies on inapposite cases involving "unauthorized" uses of *legally acquired* material. That a use is *unauthorized*, however, is quite different than where the work's very acquisition is *illegal*. Indeed, every prima facie infringement claim necessarily requires the use to be unauthorized. *See* 9th Cir. Model Jury Inst. 17.5 ("Anyone who copies original expression from a copyrighted work during

the term of the copyright *without the owner's permission* infringes the copyright.") (emphasis added). Unauthorized uses of legally acquired works—as existed in *Google Books*—are far afield from the facts here.

Meta's attempt to distinguish *Harper & Row,* the Supreme Court's seminal fair use case, falls flat. Meta Br. at 20. There, the defendant "knowingly exploited a *purloined* manuscript" and then "supplant[ed] the copyright holder's commercially valuable right of first publication" without "even the fiction of consent as justification." *Harper & Row Publishers, Inc. v. Nation Enterprises*, 471 U.S. 539, 562-63 (1985) (emphasis added). That theft without the copyright owner's consent was found *not* to be fair use. Here, Meta similarly purloined Plaintiffs' works without consent—and on a scale dramatically greater than the 13% of a single manuscript that was excerpted in *Harper & Row*.

Meta also fails in attacking the holding in *Atari Games Corp. v. Nintendo of America, Inc.*, 975 F.2d 832 (Fed. Cir. 1992) that "an individual must possess an authorized copy of a literary work" to "invoke the fair use exception," as supposedly outdated and superseded. Meta Br. at 21. Not one of Meta's "distinguishing" cases involve an unauthorized initial copy like the stolen source code in *Atari*, and none even mention the legality of the initial acquisition. In *Warhol*, the artist had previously *granted a license* to Vanity Fair to use her Prince photo as an "artist reference for an illustration," which then led to the creation of Orange Prince. *Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith*, 598 U.S. 508, 508 (2023). Similarly, *Oracle* involved no illegal piracy—the declaring code was acquired legally. *Google LLC v. Oracle Am., Inc.*, 593 U.S. 1, 13 (2021). And *Campbell* likewise does not suggest 2 Live Crew illegally pirated the song "Oh, Pretty Woman" before parodying it. *Campbell v. Acuff-Rose Music, Inc.*, 510 U.S. 569, 572–73 (1994).

Meta's amici fare no better. One attempts to limit *Harper & Row* through dicta from an unpublished, out-of-circuit district court order. Dkt. 516 at 11 ("Other courts have read this holding to require, for example, obtaining a work through 'breach of confidence or deception.'") (quoting *Am. Inst. of Physics v. Winstead PC*, 2013 WL 6242843, at *12 (N.D. Tex. Dec. 3, 2013)). But that court simply rejected the plaintiffs' argument that an initial copy of a work must categorically

be obtained "in 'good faith,'" whereas the initial copies in *Harper & Row* (and other cases) were, like here, "obtained in improper ways for known impermissible purposes" such as "st[ealing] the entire potential market for [the publisher's] materials." 2013 WL 6242843, at *12. Meta makes a similar argument, proclaiming in a footnote that it just "copied third party datasets . . . from publicly available websites; there is no evidence that it lied to obtain those copies." Meta Br. n.13. Meta's "publicly available" euphemism is akin to arguing that a book in a bookstore is publicly available because anybody can walk in during business hours, take it off the shelf, and walk out without paying for it. That's illegal, even if you do not lie to the bookstore attendant while doing it. Moreover, Meta went to great lengths to conceal its conduct, *e.g.*, Ex. 61 at -245 ("In **no case** would we disclose publicly that we had trained on libgen") (emphasis in original), including by using a concealed VPN rather than Meta IP addresses so internet service providers could not track its pirating activities. And to the extent amici argue *Perfect 10, Inc. v. Amazon.com, Inc.* "implicitly reject[s]" *Atari*, amici neglect to mention that "*Perfect 10* accepted the 'general rule' from *Atari* that 'a party claiming fair use must act in a manner generally compatible with principles of good faith and fair dealing.'" *Triller Fight Club II LLC v. H3 Podcast*, 2023 WL 11877604, at *8 (C.D. Cal. Sept. 15, 2023) (quoting *Perfect 10*, 508 F.3d 1146, 1164 n.8 (9th Cir. 2007)). Meta plainly did not.

B. Plaintiffs Need Not Prove Actual Distribution of Their Books To Prevail On Summary Judgment.

Meta attempts to raise a dispute about whether it actually uploaded Plaintiffs' Books, but that is of no moment. "As the Ninth Circuit has explained, . . . '[b]oth uploading and downloading copyrighted material are infringing acts.'" *IMAPizza, LLC v. At Pizza Ltd.*, 334 F. Supp. 3d 95, 119 (D.D.C. 2018) (quoting *Columbia Pictures Indus., Inc. v. Fung*, 710 F.3d 1020, 1034 (9th Cir. 2013)). "Therefore, where both the uploading and the downloading are unlawful (as in the many cases of illegal file-sharing that have come before the federal courts), 'Plaintiffs need only show that United States users either uploaded or downloaded copyrighted works; Plaintiffs need not show that a particular file was both uploaded and downloaded entirely within the United States.'"

Id. (quoting *Columbia Pictures Indus., Inc. v. Fung*, 2009 WL 6355911, at *8 (C.D. Cal. Dec. 21, 2009)). Applying this reasoning here, Plaintiffs have established Meta's P2P sharing of copyrighted material resulted in mass infringement from Meta's downloading alone. Further, Meta offers no facts to dispute the overwhelming probability that it also uploaded Plaintiffs' works, *see* Ex. 71, ¶¶ 20-30 & Table 1,[4] *or* at minimum, made them available to peers such that they are properly deemed distributed.

1. Meta "destroyed" evidence that might have identified content it distributed, but aggregated data nonetheless proves massive uploading of copyrighted material.

Fundamentally, Meta seeks to preclude partial summary judgment by pointing to an absence of book-by-book uploading data. While Meta maintained detailed indices of the copyrighted works it downloaded from Library Genesis ("LibGen"), Z-Library ("Z-Lib"), and Internet Archive ("IA"), *see* Exs. 94-96, it did not maintain or preserve any logs showing *which pieces* of *which works* Meta uploaded to others. Meta failed to preserve logs related to outbound data by letting its "devserver" logs (to which Meta torrented a copy of LibGen in early 2023) be "overridden" long ***after*** this lawsuit was filed on July 7, 2023. Ex. 98, King 30(b)(6) Tr. 117:1-122:20 (noting "no logs had been located" because "it was outside the [one-year] retention period," and while it is "possible to prevent such overwriting from taking place," Meta did not do so); Ex. 99, Meta_Kadrey_00112111, at -213 (Bashlykov discussing syncing LibGen data from his devservers in May 2023). Additionally, on the AWS EC2 computers that Meta used for its *2024* torrenting via Anna's Archive, the "instances" (or virtual computers) were also "destroyed" well after this case began in July 2023. Ex. 98, 196:7-200:15. Meta's aggregated AWS billing data at least enabled the parties to reconstruct the massive amount of pirated content that Meta uploaded in *2024*; no such reconstruction will ever be possible with respect to Meta's 2023 LibGen

[4] Meta objects to Dr. Choffnes's report as an "unsworn expert report" "not admissible" on summary judgment. Meta Br. at 34-35 & n.19. Courts universally allow a party to address any such objection, however, by "proffering the sworn deposition or declaration of the expert." *E.g.*, *Liebling v. Novartis Pharmaceuticals Corp.*, 2014 WL 12576619, at *2 (C.D. Cal. Mar. 24, 2014) (collecting cases). Plaintiffs thus proffer Exhibit 100, the sworn deposition of Dr. Choffnes taken March 28, 2025, where he authenticated his expert report on pages 125-126, among other places.

torrenting. These acts of "overriding" help explain why Meta's logs are absent from discovery and why both parties' torrenting experts can speak only in terms of probabilities and aggregated billing data.[5]

Nonetheless, undisputed evidence establishes that Meta uploaded at least *some* components of Plaintiffs' Books. First, it is undisputed that Meta torrented at least 134 terabytes of data from known pirated databases in April to June 2024 alone, and it is now undisputed that Meta also torrented over two million works from LibGen Fiction in 2022. Ex. 98, 46:2-47:9 (admitting LibGen was acquired via torrent in 2022); Butterick Decl. ¶¶ 2-6 (showing quantity of 2022 LibGen documents). Second, it is undisputed that those massive repositories of pirated works included at least one copy of every Book asserted in this case. Ex. 72, Table 2 & App'x C. Finally, Meta offers no evidence that its engineers did anything to modify BitTorrent's default "leeching" protocol that enables uploading simultaneously to downloading. Pltfs' Br. at 14. And it is now undisputed that Meta distributed up to ***30%*** of what it downloaded, at least with respect to the over 134 TB it torrented in 2024, Dkt. 492 at ¶ 29. Dr. Choffnes analyzed this evidence and concluded that the probability of Meta uploading at least *some* portion of *one* of Plaintiffs' Books was well over 99%. Ex. 71, ¶¶ 20-30 & Table 1.

Statistical evidence is often used in P2P infringement cases because definitive facts about illegal file sharing are difficult to obtain—by design. For that reason, courts have granted summary judgment to plaintiffs in such cases based on statistical evidence even where the underlying assumptions are disputed. *E.g.*, *Columbia Pictures*, 710 F.3d at 1034 (affirming grant of summary judgment to plaintiff on liability where proof of widespread infringement was established via expert analysis showing that "between 90 and 96% of the content associated with the torrent files available on [defendant's] websites are for 'confirmed or highly likely copyright infringing' material," and holding that "even giving [defendant] the benefit of all doubts by tripling the margins of error in the expert's reports, Columbia would still have such overwhelming evidence

[5] Plaintiffs only learned of this spoliation during the April 2, 2025 deposition of Meta's 30(b)(6) representative, and reserve the right to seek relief under Fed. R. Civ. P. 37 at the appropriate time.

that any reasonable jury would have to conclude that the vastly predominant use of [defendant's] services has been to infringe copyrights"). This case warrants that same disposition.

2. There is no dispute that Meta made Plaintiffs' Books *available* for sharing with other pirates.

It is beyond dispute that Meta's actions resulted in the distribution of millions of copyrighted works. And while no evidence shows precisely which works Meta distributed, it is likewise beyond dispute that Meta *made available* each of Plaintiffs' Books. These facts constitute additional "uses," entirely divorced from LLM training, that are categorically unfair.

Even if what Meta actually shared with other digital pirates is relevant, courts routinely accept evidence that the defendant "made available" copyrighted material in lieu of requiring proof actual sharing. In *Hotaling v. Church of Jesus Christ of Latter-Day Saints*, the Fourth Circuit rejected the notion that "the evidence would need to show that a member of the public accepted" "an offer to distribute the work" to "establish distribution." 118 F.3d 199, 203 (4th Cir. 1997). To the contrary, once an infringing party "has completed all the steps necessary for distribution to the public," that work is deemed distributed, and for good reason: "[w]ere this not to be considered distribution . . . a copyright holder would be prejudiced by a library that does not keep records of public use, and the library would unjustly profit by its own omission." *Id.* In-circuit cases have expressly acknowledged *Hotaling*'s direct applicability to the P2P file sharing context.[6] *E.g.*, *SA Music, LLC v. Amazon.com, Inc.*, 2020 WL 3128534, at *4 (W.D. Wash. June 12, 2020) (describing *Hotaling*'s facts as "analogous to ***making a work available*** to the public through a file-sharing network") (emphasis added). Indeed, once the lens is narrowed to the P2P context, "according to . . . [the] majority of cases analyzing the issue, the distribution requirement . . . is satisfied when a copyrighted work is ***made available*** for downloading through a file-sharing

[6] The Ninth Circuit cases that decline to follow *Hotaling* involve factually distinct scenarios where the infringer did not make entire copyrighted works available for free. *E.g.*, *Perfect 10*, 508 F.3d at 1162-63 ("Though Google indexes these images, it does not have a collection of stored full-size images it makes available to the public. Google therefore cannot be deemed to distribute copies of these images under the reasoning of *Napster* or *Hotaling*"). That is decidedly not the case here, where there is undisputed evidence Meta *did* distribute massive amounts of copyrighted material.

network – as one might do on a P2P network such as Napster." *Id.* (emphasis added).[7] And as Meta did here.

Meta responds that because of its destruction of its own torrenting logs, Meta cannot be deemed to have distributed any specific work, even though Meta uploaded a massive amount of copyrighted material in the aggregate. But Meta, which bears the burden to "identify with reasonable particularity the evidence that precludes summary judgment," *Keenan v. Allan*, 91 F.3d 1275, 1279 (9th Cir. 1996), offers no evidence here other than a lack of book-by-book information about its torrenting activity—the same evidence Plaintiffs might have obtained had Meta preserved it. Because there is no dispute that Meta uploaded massive amounts of copyrighted data, and there is no dispute that Meta made Plaintiffs' Books available for upload, partial summary judgment is warranted on the basis of Meta's distribution, as well as Meta's copying, of the Books through P2P networks.

II. META'S MULTIPLE OTHER INFRINGEMENTS WERE NOT FAIR USE

Nowhere in Meta's motion does it acknowledge that "the Supreme Court and [the Ninth] circuit have unequivocally placed the burden of proof on the proponent of the affirmative defense of fair use." *Dr. Seuss Enters. v. ComicMix LLC*, 983 F.3d 443, 459 (9th Cir. 2020). Meta thus "bears the ***heavy*** burden of showing there are no genuine issues of material fact about whether its copying was fair." *Nat'l Fire Prot. Ass'n v. UpCodes, Inc.,* 753 F. Supp. 3d 933, 954 (C.D. Cal. 2024) (quoting *Ambat v. City and Cnty. of S.F.*, 757 F.3d 1017, 1031 (9th Cir. 2014)) (emphasis added).

---

[7] *See also A&M Recs. v. Napster*, 239 F.3d 1004, 1019 (9th Cir. 2001) ("it is obvious that once a user lists a copy of music he already owns on the Napster system in order to access the music from another location, the song ***becomes available*** to millions of other individuals") (emphasis added) (citing *A&M Records, Inc. v. Napster, Inc.*, 114 F. Supp. 2d 896, 913 (N.D. Cal. 2000) ("a Napster user who downloads a copy of a song to her hard drive may ***make that song available*** to millions of other individuals, even if she eventually chooses to purchase the CD. So-called sampling on Napster may quickly facilitate unauthorized distribution at an exponential rate")) (emphasis added); *Columbia Pictures*, 710 F.3d at 1033 ("one can infringe a copyright through culpable actions resulting in the impermissible reproduction of copyrighted expression, whether those actions involve ***making available*** a device or product or providing some service used in accomplishing the infringement") (emphasis added).

A. Meta's Motion Ignores a Plethora of Copyrighted Book "Uses."

It is undisputed that Meta made copies of Plaintiffs' Books for multiple purposes, starting with its online piracy via torrenting and direct downloads to obtain them for free, then continuing through, *inter alia*, various stages of the LLM training process. *See, e.g.*, Ex. 101, Meta_Kadrey_00065631; Ex. 56; Ex. 102, Bashlykov 30(b)(1) Tr. 62:15-63:17, 66:5-7, 82:8-16, 147:14-148:2, 156:15-21, 225:5-226:3, 228:12-230:8 (discussing making additional copies of LibGen for various uses); Ex. 103, Bashlykov 30(b)(6) Tr. 67:9-18, 68:14-70:7, 73:17-74:7, 80:4-10.[8] Yet, Meta advances a *single* fair use argument predicated on LLM training—a defense that could only apply to copies it *actually used* to train its LLMs.

The Court should reject Meta's attempt to turn its "fair use" defense as to a single use into an umbrella defense that excuses all copying for any purpose. Different reproductions require separate analyses, and *all* uses of infringing material must be "fair use" to overcome liability. *E.g.*, *Warhol*, 598 U.S. at 533 ("the same copying may be fair when used for one purpose but not another"); *Campbell*, 510 U.S. at 585 (counseling that different outcomes may result from use of a copyrighted work to advertise a product, even in a parody, versus the sale of a parody of that same copyrighted work); *Walker v. Univ. Books, Inc.*, 602 F.2d 859, 864 (9th Cir. 1979) (that an "infringing copy" of a work is "an inchoate representation of some final product to be marketed commercially does not in itself negate the possibility of infringement"; that "the blueprints themselves were never sold for profit" does not "eliminate the possibility of an award of statutory damages for infringement under the Act"). Here, Meta made numerous copies of pirated datasets of copyrighted works from different shadow libraries at different times, which it stored in different

[8] To date, Plaintiffs have been unable to discover all of the uses that Meta has made of Plaintiffs' Books and other copyrighted works. Every time Plaintiffs look somewhere new, more illicit copies appear. Just last week, Plaintiffs discovered source code showing that Meta began downloading pirated data via blockchain, using the so-called InterPlanetary File System ("IPFS"), where many shadow libraries taken down from the public internet by the FBI, court order, or other means still reside. Meta employees had previously discussed the availability of LibGen via IPFS, *e.g.*, Ex. 56 at -861.00003, but this source code revealed the first evidence of Meta starting to actually use the IPFS. In light of the lack of evidence of the full extent of Meta's copying, which is squarely within Meta's control, the appropriate inference is that Meta made multiple copies of the pirated works, some of which were used for LLM training, but others of which were not. Fed. R. Civ. P. 56(d).

locations for different uses and purposes. The following copies are indisputably not fair use:

Fall 2022: In 2022—as Meta employees were in talks with at least one publisher for AI training data, Ex. 32—Meta torrented over two million pirated books from LibGen. The documents reflecting that torrenting appear to contain 204 copies of Plaintiffs' Books. Butterick Decl. at ¶¶ 2-6; Ex. A (summary table). Meta never used this dataset to train any Llama model.[9] Instead, the sole use and purpose of this dataset was to determine "if there [wa]s value" in copyrighted books as training data. If there was, Meta intended to "setup proper licensing agreement[s]"—something Meta never did. Ex 32, Meta_Kadrey_00218170; *see also* Ex. 104, Meta_Kadrey_00218907, at -909 (stating Meta would "use libgen to benchmark the performances based on the amount of data, and use that to decide which licenses to buy for the actual model").

At the time, the head of Meta's AI group wondered if it was "problematic" to use LibGen for this purpose. Ex. 32. Her concerns were well-founded. Using LibGen to source pirated market substitutes of high value AI training data is quintessentially not fair use. The record shows the following: (1) Meta torrented millions of works of fiction from LibGen even as employees internally protested using an "illegal pirated website" for training data, Ex. 18 at -730; (2) Meta used this dataset to determine whether the data was valuable enough to license, Ex. 105, Meta_Kadrey_00231286; and (3) Meta never used this dataset to train any Llama model because, at the time, it concluded the risks far outweighed the rewards. Ex. 35 ("The team was initially evaluating the usage of copyrighted work such as libgen fiction . . . Results obtained in December show that the upside from using these is negligible, while the legal limitations in order to train a LLM for widespread internal use are significant."). In other words, Meta infringed millions of copyrights—including Plaintiffs'—but ultimately decided to abandon this dataset without "using" it to train any of its LLMs.

[9] *See* Meta Answer, Dkt. 485, ¶¶ 3, 35, 98 (admitting Meta "made copies" of portions of LibGen and other shadow libraries, and used that text data, which included books, for not just training but also LLM research and evaluation, including "assess[ing] the general knowledge and expressive abilities of models, and as a means of testing LLMs' memorization").

Spring 2023: Meta's decision not to use pirated works from LibGen for model training was short-lived. Several months after resolving not to use that data, Meta made the business decision to reverse course. However, Meta's internal culture of competition meant that different engineering teams refused to share training data. While Meta engineer Nikolay Bashlykov tried to gain access to the 2022 LibGen dataset, those efforts fell short. Ex. 106, Meta_Kadrey_00101706 (Bashlykov requesting LibGen access; no response from Lample). He then downloaded and torrented a second copy of LibGen. Ex. 70. Yet Meta's first use of that new LibGen copy again was *not* to train an LLM. Rather, it was used for the express purpose of determining whether Meta had any reason to license copyrighted works beyond what LibGen already contained. Ex. 56 at -861.00018; Ex. 58 ("Do we still want to buy [redacted] despite the similarity [with LibGen] that Nikolay found . . .?"). Contrary to the Bashlykov declaration that Meta filed, which claims that "[t]he selection process did not involve reviewing individual book titles or articles," Dkt 494 at ¶ 6, the record shows Meta engineers were instructed to do exactly that. They directly cross-referenced the works available in LibGen against catalogues of books offered by literary publishers to determine "to what extent libgen data already has papers/textbooks we could purchase from [redacted]" Ex, 57, Meta_Kadrey_00160779. Specifically, Meta compared titles for potential licensing to the titles available in LibGen. Ex. 75 (spreadsheet of titles from [redacted] and link for copy on LibGen if it exists with count of "match," "miss" and "share" of books compared to LibGen). Bashlykov testified he used one LibGen copy for a "book-matching" exercise comparing "results in LibGen versus entries contained within a publisher's catalog," before ultimately determining that Meta did not need "to proceed with [redacted]" "[b]ecause 90 percent of [redacted] was in LibGen already." Ex. 102, 94:6-11, 165:6-18; Ex. 56 at -861.00018; *see also* Ex. 107, Meta_Kadrey_00233772 ("[redacted] Backpedal. LibGen: Could be coming").

Summer 2024: Meta then engaged in a nearly identical "use" of pirated database Anna's Archive. In Summer of 2024, Meta fully torrented Anna's Archive, which included pirated databases Z-Lib, IA, and LibGen once again. *See* Ex. 108, Meta_Kadrey_00237719, at -720 (distinguishing between "Libgen 2023" and Libgen 2024" as separate copies in Llama's training

data mix). Once Anna's Archive was fully ingested, however, Meta realized it could not acquire all copyrighted works via piracy alone. Ex. 97 at -195 ("Anna's Archive shows that we have currently only collected 5% of the world book content."). To acquire the rest, Meta finally concluded it would need to license text data—including books. Meta developed a straightforward strategy: "cross reference books we have collected from Anna's Archive" against "all the books in the world" to "manually evaluate and hypothesize whether the missing data corpus from these publishers would be valuable to acquire." *Id.* at -196. Meta emphasized that it "**will only** go after the publishing companies with the largest delta of missing content" between their offerings and what Meta "collected from Anna's Archive & spidermate."[10] *Id.* at -198 (emphasis in original).

Meta's acquisition of these pirated datasets was a chaotic free-for-all. Meta employees downloaded several copies of pirated datasets, only some of which ever made it into an LLM for training. *See* Ex. 109, Nayak 30(b)(6) Tr. 66:6-9 ("Meta is a large company, and I could imagine that one researcher could have downloaded [LibGen] and then another researcher could have downloaded it without knowing that it existed."). Meta does not argue—and there is no evidence to indicate—that any of Plaintiffs' Books were transformed in any way during the many times it copied them from LibGen and Anna's Archive. At a minimum, the initial download of copyrighted works "does not transform the copyrighted work." *Napster*, 239 F.3d at 1014; *see Am. Geophysical Union v. Texaco, Inc.*, 60 F.3d 913, 923 (2d Cir. 1994) ("Texaco's making of copies cannot properly be regarded as a transformative use of the copyrighted material"). And there is nothing transformative about Meta's use of copyrighted works in LibGen and Anna's Archive to analyze whether pirated datasets are adequate market substitutes for paying to license *those same books*—after all, substitution is "copyright's bête noire." *Warhol*, 598 U.S. at 528.

B. For the Specific "Use" of LLM Training, There Are Factual Disputes for Trial.

Even for the lone "use" that Meta *does* address in its Motion—copying Plaintiffs' Books "to train Llama," *see* Meta Br. at 13, there exist substantial factual disputes that foreclose summary

[10] Spidermate is Meta's web crawling application. It scrapes a vast set of web data, including copyrighted content, used by Meta for LLM training, entirely separate from Meta's torrenting. *E.g.*, Ex. 108 (showing Spidermate in training data mix); Ex. 143 (explaining Spidermate).

judgment. Certain fair-use factors lean heavily in Plaintiffs' favor, and all four factors must be balanced against each other—a "fact specific inquiry for which summary judgment is ill-suited," particularly where the disputed use lacks any directly applicable precedent, like here. *See Ass'n of Am. Medical Colleges v. Cuomo*, 928 F.2d 519, 524 (2d Cir. 1991); *In re DMCA*, 608 F. Supp. 3d at 879 (Chhabria, J.) (noting that apart from "obvious" cases like "downloading and distributing copyrighted music via peer-to-peer systems," "[t]he fair use analysis is fact intensive").

1. Factor One: The Purpose and Character of the Use

The first factor is the purpose and character of the use. 17 U.S.C. § 107(1). In weighing this factor, courts consider whether the use was commercial and whether it was transformative. *Warhol*, 598 U.S. at 529-31. Commerciality and transformativeness are two related but distinct inquiries. "[T]he degree of difference [between the original work and the secondary use] must be balanced against the commercial nature of the use." *Id.* at 532. Importantly, a "[t]ransformative use is neither absolutely necessary *nor sufficient* for a finding of fair use." *Barcroft Media, Ltd. v. Coed Media Grp., LLC*, 297 F. Supp. 3d 339, 351 (S.D.N.Y. 2017) (quotations omitted) (emphasis added).

a. Llama is a commercial product.

All of Meta's uses of Llama are highly commercial and profit-seeking. Meta soft-pedals this, euphemizing that "it hopes one day to recoup its significant investment" in Llama. Meta Br. at 18. The bare truth is that Meta forecasts its GenAI products will generate between **$460 billion** and **$1.4 trillion** of total revenue by 2035. Ex. 8 at -020. Meta is not a charity or 501(c)(3): it is one of the world's largest public, for-profit companies, and it developed its series of LLMs as part of a calculated bet to generate potentially massive profits from GenAI. Pltfs' Br. at 3-4.

Further, Meta's cited authority does not support its contention that "Llama is used for both commercial and non-commercial purposes." Meta Br. at 18. For the single use Meta addresses—LLM training—Meta depends entirely on the argument that Llama is "transformative" as to the copyrighted books at issue, without balancing any supposed transformativeness with commercialism. *See Warhol*, 598 U.S. at 525 ("new expression . . . is not, without more,

dispositive of" factor one, and transformativeness "must be weighed against other considerations, like commercialism"). Indeed, Meta's commercial use "loom[s] larger" in a case like this, where Meta's use has no justification that it is somehow commenting on or otherwise message-targeting the almost innumerable works it has infringed. *Warhol*, 598 U.S. at 547 (quoting *Campbell*, 510 U.S. at 580) (alteration in original); *see Dr. Seuss*, 983 F.3d at 452-53; *see also Google Books,* 804 F.3d at 215 ("A secondary author is not necessarily at liberty to make wholesale takings of the original author's expression merely because of how well the original author's expression would convey the secondary author's different message."). Instead, Meta, with its tens of billions of dollars of annual investment into GenAI, had repeated opportunities to obtain text data to train its LLMs by compensating rightsholders but *chose* to rely on pirated substitutes to get data it wanted for free. This, too, is evidence of a commercial purpose.

b. Meta's use of copyrighted works to train LLMs is not transformative.

Unlike Google's use of copyrighted works to build a digital index in *Google Books*, Meta uses copyrighted works to create a commercial product capable of writing books or mimicking the style of writers on whose works Meta has trained. That is why Meta cares about data diversity: each type of data source affects what Meta's Llama models can output. Record evidence shows Meta intentionally designed its LLMs to emulate the writing style of specific authors. Meta's training on copyrighted works does not advance any traditionally protected categories of transformative use, and the mechanics of LLM technology are designed to mimic copyrighted works rather than transform them. At minimum, this is a factual dispute that belongs with the jury.

i. Meta's LLMs leverage copyrighted works to copy the protected expression contained within them.

Meta claims that its use of Plaintiffs' Books is transformative because "Llama is nothing like a book; it is not meant to be read." Meta Br. at 3. According to Meta, because books are little more than statistical training fodder in this use context, the purpose of LLM training is entirely new. But it is not that simple—not even close. Books are used as LLM training data *because they are books*, not because they are mere amalgamations of words on a page. Llama might not

regurgitate entire novels verbatim due to Meta's post-training copyright "mitigations," but when a Llama user asks the model to write a work of fiction in the style of Richard Kadrey, Llama will do it. That output can only occur because the model was extensively trained on books, including Kadrey's books. Llama does not *transform* Kadrey's works, it *mimics* them as closely as possible.

Record evidence reveals that Meta employees used books for far more than their statistical patterns. Employees read works Meta copied for training data and went to great lengths to train Llama models to produce outputs aligning with specific book authors. Meta employees routinely cited "creativity" as the reason it sought books such as fiction novels. *E.g.*, Ex. 110, Meta_Kadrey_00080277, at -279. Meta, even from the earliest days of its GenAI program, built Llama to directly compete with writers as a means of creating prose and narrative text. *E.g.*, Ex. 111, Meta_Kadrey_00074217. Indeed, Meta was developing its LLMs to mimic specific authors. For example, Meta produced fine-tuning data (*i.e.*, post-training data) intended to replicate the writing style of named Plaintiff Junot Diaz. Ex. 112, Meta_Kadrey_00019627; Ex. 113, Meta_Kadrey_00028015. Moreover, several senior Meta AI scientists, including some of the lead developers of Llama, had a lengthy discussion regarding their work to train Meta's LLMs to write "in the style of" certain book authors, including Kurt Vonnegut and William Faulkner—both of whose works are still under copyright. Ex. 114, Meta_Kadrey_CT_00086196, at -206, -237, -238.

Mimicking copyrighted expression—a quintessential non-transformative use—goes to the very heart of how LLMs operate. Plaintiffs' expert Emily Bender opines that LLMs merely "model[] patterns in their training data," so it is inaccurate to claim they "'understand[]' language or otherwise hav[e] access to 'meaning.'" Ex. 115, Bender Report, ¶ 27. LLMs are, in Dr. Bender's phrase, "stochastic parrots": devices that "repeat[] something without understanding it." Ex. 116, Bender Tr. 76:4-19. Unsurprisingly, feeding copyrighted material into LLMs results in parroted outputs derived from that very same copyrighted material. In fact, LLMs cannot generate high-quality responses *unless* they are trained on comparable material. Ex. 115 ¶ 22 ("an LLM being used to synthesize text will only be able to output text that looks like literary fiction if its training data includes sufficient examples of literary fiction"). For that reason, training an LLM on books

results in ***copying the very heart of the protected expression within them***. *See id.* ¶ 90 ("The import of a text lies not just in its length, but in the word choice and grammatical structures it includes. These are the patterns that the LLMs are designed to represent and serve as the basis for their functionality in outputting text that mimics all of the genres in their training data.").

Thus, Meta does not merely use books as statistical training inputs that are somehow agnostic to expressive content. Meta relied on books due to that content—what humans wrote—and then trained LLMs to emulate protected expression from the very authors of those works. In that sense, Meta's use of Plaintiffs' Books as training data is not transformative because it merely amounts to a "repackaging" of those works in a different format. Meta "merely transforms *the material object* embodying the … work" into a tokenized training version. *Texaco*, 60 F.3d at 923 (italics in original). Moreover, despite Meta's claim that it "used the copies it made to develop and train Llama," Meta Br. at 36, unauthorized repackaging of copyrighted books is not inherently "integral to transformative and productive ends of scientific research." *Texaco*, 60 F.3d at 932 (Jacobs, J., dissenting). Meta therefore "cannot gain fair use insulation . . . simply because such copying is done by a company doing research." *Id.* at 924. Meta's repeated repackaging of expressive copyrighted books, even in the course of developing new technology, was not fair use.

ii. Meta's LLMs have nothing to do with any traditionally transformative uses.

Fair use prioritizes "purposes such as criticism, comment, news reporting" and other uses that "target" the work in question. 17 U.S.C. § 107. Those recognized uses "shed light on the original's depiction." *Warhol*, 598 U.S. at 547 n.21. Further, the question of transformativeness "is a matter of degree," *id.* at 532, and where a subsequent work has "no critical bearing" on the prior one such as for criticism or parody, its "'claim to fairness . . . diminishes accordingly[.]'" *Id.* at 546-47 (quoting *Campbell*, 510 U.S. at 580). Copying may be "helpful" for conveying the meaning of the subsequent work, "but that does not suffice" to establish transformativeness because helpfulness alone would not separate legitimate fair users from "would-be fair users" like "a musician who finds it helpful to sample another artist's song." *Id.* at 547-48; *Dr. Seuss*, 983

F.3d at 453 (evoking an original work while failing to ridicule or comment on it favors finding against fair use).

Meta asks the Court to find that "Llama is radically transformative" primarily because "it is an entirely new technology." Meta Br. at 15. But *what Llama actually does* dramatically weakens this claim: according to Meta, Llama "can serve as a personal tutor[,] . . . assist with creative ideation, and help users to generate business reports," and so forth. *Id.* at 3. None of those end uses have any "critical bearing" on the copyrighted works at issue. *Warhol*, 598 U.S. at 546. This lack of targeting the underlying works thus substantially weakens Meta's claim of transformativeness.

iii. Meta's use of Plaintiffs' Books is not protected intermediate copying.

Meta and its amici next argue that any copies created during the LLM training process are nevertheless protected as fair use under what they characterize as the intermediate copying "exception." Meta Br. at 17 (citing *Sony Comput. Ent., Inc. v. Connectix Corp.*, 203 F.3d 596, 599 (9th Cir. 2000)); IP Law Professors' Amicus Brief, Dkt. 509-2 at 8-12. Both briefs misconstrue the "intermediate copying" doctrine.

The Copyright Act "unambiguously encompasses and proscribes 'intermediate copying'." *Sega Enterprises Ltd. v. Accolade, Inc.*, 977 F.2d 1510, 1518 (9th Cir. 1992), *as amended* (Jan. 6, 1993) (citing *Walker,* 602 F.2d at 864 ("the fact that an allegedly infringing copy . . . may itself be only an inchoate representation of some final product . . . does not . . . negate the possibility of infringement")). In the rare instances when intermediate copying has qualified as fair use, it was the "only way to gain access to the ideas and functional elements embodied in a copyrighted computer program." *Connectix*, 203 F.3d at 602 (citing *Sega*). These preconditions, as established in *Sega*, plainly do not apply to Meta's conduct here.

First, protected "intermediate copying is generally limited to cases involving software." *Cline v. Reetz-Laiolo*, 329 F. Supp. 3d 1000, 1036 (N.D. Cal. 2018).[11] The material that Meta copied in this case is not a "copyrighted computer program," but rather constituted entire books

[11] Llama, of course, contains software. But that software is not what was unlawfully copied, unlike in *Sega* and its progeny.

and other copyrighted textual material. Nor were the copies Meta made "intermediate" at all. Meta has effectively created a *permanent* stockpile of millions of copyrighted books and other textual works to train Llama. Ex. 162, Meta RFAs (admitting Meta stores copyrighted material for training Llama Models). That stockpile persists at Meta to this day; recently produced documents show Meta continued to run tests on the pirated books it torrented from Anna's Archive as recently as ***February 2025***. Ex. 117, Meta_Kadrey_00237186 (100-page digest showing Meta running near-continuous tests on pirated datasets from March 2024 through February 2025).

Second, the *Sega* line of cases only applies "to the ideas and functional elements" in the copied work, and only when the copying is done "solely" for this purpose. *Sega*, 977 F.2d at 1527. Meta concedes that "Plaintiffs' works are creative" but says "the aspects of the works that Meta needed to extract and use to train Llama are unprotected statistical data." Meta Br. at 22. However, Meta does not—because it cannot—argue that it copied Plaintiffs' works solely for this reason, because Meta wanted and took their expressive content. *See* § II.B.2, *infra*. Undisputed evidence establishes that Meta copied copyrighted works repeatedly and nearly in their entirety. Pltfs' Br. at 20-21.

Third, *Sega* requires that the intermediate copying be "necessary in order to gain access" to the needed material. 977 F.2d at 1519. Importantly, however, the *Sega* court used the word "necessary" in the strong sense of "no other method . . . was available" to the defendant. *Id.* at 1522. Though Meta claims broadly that its "use of whole books to train Llama was necessary," it never suggests it had "no other method" of gathering equivalent data. Meta Br. at 23. On the contrary, Meta could have acquired many of the exact same copyrighted works legitimately through training-data licensing arrangements. *See* § II.B.4, *infra*. Thus, pirating books was not "necessary" for training.

Finally, in *Sega*, the court stressed that its fair use determination relied on the fact that the final product did not contain any infringing material. But here, extensive record evidence shows regurgitation of training data is a well-known phenomenon that Meta viewed as a significant problem. *E.g.*, Ex. 118, Meta_Kadrey_00054433 ("Llama is annoyingly good at quoting books

🙃."); Ex. 119, Meta_Kadrey_00000277 (internal Meta presentation titled "Memorization in LLMs & MME [Memorization Measurement Engine]"); Ex. 120, Esiobu Tr. 62:7-15; 82:22-83:4 (observing regurgitation of text from LibGen by Llama and regurgitation even of fine-tuned Llama models); Ex. 121, Meta_Kadrey_00049649 (identifying copyright books as a memorization problem"); Ex. 122, Meta_Kadrey_00063146, at -156 ("North Star: Ensure that models do not memorize and regurgitate training data"); *id.* at -157 ("If we end up just 'obfuscating' data that we train by preventing the model from regurgitating it verbatim, that doesn't seem super ethical."). The regurgitation concerns are not limited to Meta's fact witnesses. Meta's own expert, Dr. Ungar, identified significant memorization effects for books, including Plaintiffs' Books. Ex. 123, Ungar Tr. 115:8-14 ("What I found was an average for each book, one short passage of roughly 50 tokens could, with some statistical probability, be reconstructed under these specialized circumstances designed to make it as easy as possible to reconstruct them."); Ex. 124, Lopes Rebuttal Rep., Table 4 (compiling results of Dr. Ungar's continuation experiments). In short, unlike in *Sega*, Meta's Llama models store a memory of their training data, and there is at least a genuine dispute of material fact whether they can output infringing material. *See* Ex. 125, LeCun Tr. 129:11-133:2.

c. Meta's conduct evinces a startling degree of bad faith.

"The propriety of defendant's conduct" is "relevant to the character of the use." *Harper & Row*, 471 U.S. at 562-63. In evaluating the bad-faith subfactor, the Ninth Circuit applies "the general rule that a party claiming fair use must act in a manner generally compatible with principles of good faith and fair dealing." *Perfect 10*, 508 F.3d at 1164 n.8. As one court put it, "copyright law is not concerned with a person's generally good or bad character – but rather whether that person obtained copyrighted material using egregious and inappropriate means." *Larson v. Dorland*, 693 F. Supp. 3d 59, 83 (D. Mass. 2023).

In *Harper & Row*, the Supreme Court held there was no fair use where the defendant "knowingly exploited a purloined manuscript" because that conduct demonstrated bad faith. 471 U.S. at 563, 569. Here, Meta knowingly exploited and purloined not just a single manuscript, but *tens of millions* of copyrighted works. In addition, circumventing licensing requirements generally

evidences bad faith. *See Los Angeles News Serv. v. KCAL–TV Channel 9,* 108 F.3d 1119, 1122 (9th Cir.1997) (reversing summary judgment in defendant's favor where defendant news station "obtained a copy of the tape from another station, directly copied the original, superimposed its logo on the LANS footage, and used it for the same purpose for which it would have been used had it been paid for"). Meta plainly did that here, repeatedly cross-referencing works that could be obtained via legitimate licensing opportunities against copyrighted works it pirated to see if there was any reason to pay (and never once licensing as a result). Pltfs' Br. at 9-10; Ex. 97 at -197-98.

Instead of denying that it acted in bad faith, Meta tries to downplay the importance of the bad-faith inquiry. But the Supreme Court has never held that bad faith is irrelevant to fair use, nor has it overturned the doctrine despite multiple opportunities to do so. Rather, it only suggested that *good* faith does not carry the same weight in the opposite direction. *Campbell*, 510 U.S. at 585 n.18 (expressing skepticism about the relevance of good faith with respect to fair use). Most recently, the Supreme Court observed that the bad-faith subfactor is a "factbound consideration," without disclaiming its bearing on fair use. *Oracle,* 593 U.S. at 32-33. And because the inquiry is highly factual, it is usually ill-suited for summary judgment and instead properly decided by the jury. *See Sarl Louis Feraud Int'l v. Viewfinder Inc.*, 627 F. Supp. 2d 123, 131 (S.D.N.Y. 2008) ("whether defendant acted in bad faith presents issues of fact for resolution at trial").

2. Factor Two: The Nature of the Copyrighted Work

Factor Two weighs decisively in Plaintiffs' favor. Plaintiffs' Books—which include novels, memoirs, and plays—are original expressive works that lie at the zenith of copyright protection. *See*, *e.g.*, *Campbell*, 510 U.S. at 586 (explaining "some works are closer to the core of intended copyright protection than others," and characterizing a "fictional short story," a "soon-to-be-published memoir" and "motion pictures" as highly protected works); *see also Marcus v. Rowley*, 695 F.2d 1171, 1176 (9th Cir. 1983) (holding even a recipe book was sufficiently "creative" to render this factor neutral).

Meta attempts to maneuver around settled law by arguing that books are merely statistical data. Meta Br. at 21-22. That argument is nonsensical and contrary to the record. As no party

meaningfully disputes, Meta copied long-form books precisely *because* they contain creative elements that accurately reflect human expression—the very heart of what copyright protects. *See* Meta Br. at 1 (describing Llama as "an extraordinary technology capable of providing human-like responses"). Meta viewed books as a particularly high-quality source of LLM training data for that reason. Pltfs' Br. at 5-6. Regardless of how Meta processed the data, it preserved the arrangement of the words and therefore the creative expression inherent in the works. Ex. 115 ¶¶ 19, 82, 86.

None of the cases Meta cites are on point. In *Connectix*, a case involving reverse-engineering of computer code, the Ninth Circuit stated the copied code fell "at a distance from the core [copyright-protected works] because it contains unprotected aspects that cannot be examined without copying." 203 F.3d at 603. The unprotected elements of the computer code were not embodiments of the programmer's creative expression, but instead were purely "functional" features such as the code enabling a video game cartridge's compatibility with a console. *See Sega*, 977 F.2d at 1522-23. And *Google Books* is an immediate dead end. Meta cites it for the proposition that factor two favors fair use "where published books were copied to extract information about the words they contain, not for their expression." Meta Br. at 22. Meta completely ignores the rest of the Court's analysis, namely: (1) copyright ***does*** protect an "author's manner of expressing [unprotected] facts and ideas"; (2) those plaintiffs' books were "factual," not "fiction"; and (3) "the second factor considered in isolation" did not "influence[] us one way or the other." *Id.*

3. Factor Three: The Amount and Substantiality of the Portion Used

Meta fails to acknowledge the general rule for factor three—"[w]hile wholesale copying does not preclude fair use per se, copying an entire work militates against a finding of fair use." *Napster*, 239 F.3d at 1016. Meta concedes it knowingly copied the *entirety* of millions of copyrighted works without any compensation to their authors. *See* Meta Br. 23 (admitting Meta took and used "whole books to train Llama"). Instead, Meta argues "copying an entire work is fair where reasonable or necessary to achieve the purpose of the fair use." Meta Br. at 22-23. None of Meta's cited cases employ the term "reasonable" in this way. Not only is necessity the pertinent test in assessing Factor 3, but passing it merely gets the new user a "draw" on the factor. *Elvis*

Presley Enters. v. Passport Video, 349 F.3d 622, 630 (9th Cir. 2003), *overruled on other grounds* ("[I]f the new user only copies as much as *necessary* for his or her intended use, this factor will not weigh against the new user.") (emphasis added); *Monge v. Maya Mags.*, 688 F.3d 1164, 1179 (9th Cir. 2012) (defendant "copied 100 percent of the copyrighted photos at issue," which was "far more than was necessary to corroborate its story"). Meta's cited authority supports the same application.[12] Use of the entirety of the copyrighted works is not necessary for any of Meta's uses—or, at a minimum, there are factual issues concerning the necessity of Meta's wholesale exploitation.[13]

Additionally, Meta relies heavily on a comparison between the way the search function in *Google Books* gave users access to a limited portion of copied works and the way Meta's use purportedly "does not make any significant portion of the texts available to Llama users." Meta Br. at 23-24. That *Google Books* analysis, however, depended on finding that the search function's limited outputs had no significant effect on the copyrighted works' performance in *the sale of books*, which was the only market at issue. 804 F.3d at 222-23; *see* Meta Br. at 22:23-26 ("what matter[ed]" in *Google Books* was "'the amount and substantiality of what is thereby made

[12] *Kelly v. Arriba Soft Corp*, 336 F.3d 811, 820-21 (9th Cir. 2003) ("If the secondary user only copies as much as is ***necessary*** for his or her intended use, then this factor will not weigh against him or her.") (emphasis added); *Perfect 10*, 508 F.3d at 1167 (applying the same "necessary" analysis); *Authors Guild, Inc. v. HathiTrust*, 755 F.3d 87, 98 (2d Cir. 2014) ("***The crux of the inquiry is whether 'no more was taken than necessary***.'") (emphasis added) (quoting *Campbell*, 510 U.S. at 589) (emphasis added); *Google Books*, 804 F.3d at 221 (noting *Hathitrust*'s holding that wholesale use was found to be "reasonably ***necessary***" for the relevant use, and finding Google's wholesale book use was "***literally necessary***" to achieve Google's purpose of "advising searchers reliably whether their searched term appears in a book (or how many times)") (emphasis added); *Bill Graham Archives v. Dorling Kindersley Ltd.*, 448 F.3d 605, 613 (2d Cir. 2006) ("[C]ourts have concluded that [wholesale] copying does not necessarily weigh against fair use because **copying the entirety of a work is sometimes *necessary*** to make a fair use of the image.") (emphasis added).

[13] As in the rest of Meta's Brief, Meta's argument concerning Factor 3 is limited only to LLM training. Meta's initial use—copying the entirety of copyrighted books from known pirated sites—cannot be validly justified, and Meta does not try. *E.g.*, *Hachette,* 115 F.4th at 190 ("[IA] copies the Works in full and makes those copies available to the public in their entirety. . . . At least in this context, ***it is difficult to compete with free***.") (emphasis added); *see also Glacier Films*, 896 at 1043 ("downloading or uploading of the copyrighted work" via a P2P network cannot possibly be "permitted by the doctrine of fair use").

accessible to a public for which it may serve as a competing substitute") (emphasis in original). Factor four, *infra*, extensively discusses how Meta's use of the entirety of Plaintiffs' Books harmed them in multiple relevant markets.

Further, with respect to Llama training, Meta provides *zero* record evidence supporting its contention that using "whole books to train Llama was *necessary* for" Llama training. *See* Meta Br. at 23 (emphasis added); *id.* ("Llama's utility depends on copying whole books (and many other data sources)"). The exclusive support Meta provides for that grandiose claim is a quote from Plaintiffs' opening brief that says absolutely nothing about the reasonableness—much less the necessity—of using the *entirety* of copyrighted works, and instead only describes books as "high-quality text data" that are "uniquely valuable for developing longer-context windows." *See id.* at 23. Meta thus fails to cite any on-point support for the marginal utility of using the *entirety* of copyrighted books or other literary works, as opposed to any smaller portion of those works. *See HathiTrust*, 755 F.3d at 98; *Campbell*, 510 U.S. at 586-87; *Google Books*, 804 F.3d at 221. Factor 3 requires a fact-intensive and context-specific analysis of whether wholesale copying was necessary for any of Meta's uses, including but not limited to Llama training. *See generally Monge*, 688 F.3d at 1178-80; *Campbell*, 510 U.S. at 586-89. At the very least, Meta must present more than *zero* on-point record evidence to satisfy its heavy burden on summary judgment.

4. Factor Four: The Effect of the Use On the Potential Market for or Value of the Copyrighted Work.

The fourth factor is concerned with the threat of market substitution, expressly requiring consideration of "the effect of the use upon the potential market for or value of the copyrighted work." 17 U.S.C. § 107(4). At issue here are the fact-bound, use-specific questions of whether Meta's conduct harms multiple markets for Plaintiffs' Books or, if the conduct were "unrestricted and widespread," the potential markets for the books. *McGucken v. Pub Ocean Ltd.*, 42 F.4th 1149, 1163 (9th Cir. 2022).

Meta claims "there was no market to license Plaintiffs' works for LLM training"; "there *still* is not"; and "no such market is likely to develop." Meta Br. at 26. These statements are breathtaking in light of the factual record. From the very beginning of Meta's LLM program, Meta

believed that licensing would be required to fulfill its training data needs. Overwhelming evidence shows the existence of an already robust and growing licensing market for LLM training. Further evidence shows the existence of the longstanding market for book sales—at least 692 of which were denied to Plaintiffs by Meta's use of pirated market substitutes. Meta cannot meet its heavy burden to *disprove* market harm by simply pretending that a rapidly growing market does not exist. *McGucken*, 42 F.4th at 1163 (at summary judgment, the burden is on "the proponent of the affirmative defense of fair use [to] bring forward favorable evidence about relevant markets").

a. Meta's conduct harms the evolving market for books as AI training data.

"[A] copyright holder is entitled to demand a royalty for licensing others to use its copyrighted work, and . . . the impact on potential licensing revenues is a proper subject for consideration in assessing the fourth factor." *Hachette Book Grp., Inc. v. Internet Archive*, 115 F.4th 163, 192 (2d Cir. 2024). Meta made hundreds of copies of Plaintiffs' Books, including for use in training LLMs. This infringing use is the same one that motivates Meta and other LLM companies to license books as training data. Meta explicitly treats its repositories of pirated books as direct substitutes for licensed versions. *See* Ex. 97, at -197-98. Courts universally recognize that this sort of direct substitution inflicts a cognizable harm on the licensing market for a copyright holder's works. *See*, *e.g., McGucken*, 42 F.4th at 1163; *Monge*, 688 F.3d at 1181; *Napster*, 239 F.3d at 1016; *Texaco*, 60 F.3d at 927 (finding market harm where defendant made complete copies of individual journal articles for the same use that they are licensed for); *On Davis v. The Gap, Inc.*, 246 F.3d 152, 176 (2d Cir. 2001), *as amended* (May 15, 2001) (finding fourth factor weighed against fair use where defendant "avoided paying 'the customary price' [plaintiff] was entitled to charge for the use of his design" by taking the design for free) (quotation marks omitted).

The demand for AI training data has already been recognized as a potentially exploitable market for copyright holders. *Thomson Reuters Enter. Ctr. GMBH v. Ross Intel. Inc.*, 2025 WL 458520, at *9 (D. Del. Feb. 11, 2025) (recognizing potential market for copyrighted text as "data to train legal AIs") (Bibas, J.) (on appeal); Ex. 126, Spulber Opening Report, ¶ 23 ("[T]here is a well-established, large, and rapidly growing market for licenses to train LLMs on books and other

protectable works."). And as companies increasingly enter into training data licenses involving copyrighted literary works, the existence of an actual market for Plaintiffs' Books as training data becomes an inescapable conclusion. *See Monge*, 688 F.3d at 1181 (defendant's purchases of copyrighted pictures at issue, and prior similar pictures, "unequivocally demonstrates" a market for the copyrighted works); Ex. 97 at -197. On the supply side, many authors and Plaintiffs are willing to license their works as LLM training data. *See* Ex. 161, Plaintiffs' RFAs; *see also* Ex. 127, Silverman Tr. 203:6-7 ("I do know that I lost the sale of licensing my book to Meta"). Moreover, Meta's own conduct in this case—reflected in extensive record evidence—shows that Meta is well aware of the burgeoning market for copyrighted literary works as AI training data.

i. Meta has recognized the need to license books since the very inception of its GenAI program.

From the very beginning of Meta's GenAI program, Meta employees were already in discussions with publishers to license AI training data, including "[redacted]" Ex. 18. Licensing books for training data was *always* the end strategy for Meta's commercially available LLMs. Even when Meta first began exploring the use of pirated datasets like LibGen, it expressly disavowed using LibGen as a substitute for licensed books. Meta initially acquired its LibGen datasets to determine whether books were sufficiently valuable as training data to justify seeking out licensing deals. Ex. 18; Ex. 32. If they were, then Meta intended to "setup proper licensing agreement[s]" for this data. *Id.*; *see also* Ex. 104 at -909.

Once books proved highly valuable, Meta scoured the market to acquire collective licenses from publishers.[14] Meta inked four small deals for books in 2023. Ex. 47. The directive from Meta AI's leadership was clear: "try to get all the big guys: Harper Collins, Simon & Schuster, Macmillan, Hachette, Penguin Random House." Ex. 41. By March 2023, Meta had begun actively negotiating book licenses with [redacted]

[14] Ex. 128, Meta_Kadrey_00152235 (unsigned data license agreement with [redacted]); Ex. 129, Meta_Kadrey_00153810 (unsigned data license agreement with [redacted] dated March 17, 2022); Ex. 130, Meta_Kadrey_00153824 (unsigned data license agreement with [redacted] dated March 15, 2022); Ex. 131, Meta_Kadrey_00153924 (unsigned data license agreement with [redacted] dated June 15, 2022).

[redacted], Ex. 44, and exchanging draft data evaluation agreements.[15] Those agreements sought a "representative sample of all available eBooks across all available genres (e.g., fiction, non-fiction . . .)"—in other words, precisely the same types of works that Meta shortly thereafter decided to pirate instead. *See id.* Meta even executed a preliminary licensing agreement with [redacted]. Ex. 132, Meta_Kadrey_00146678, which it then decided to "backpedal" from after LibGen was approved for use in training commercially available models. Ex. 107. The supply of books was there—for a price—and Meta has admitted that in April 2023, "[t]here was a potential market" for licensing literary works. Ex. 133, Choudhury 30(b)(6) Tr. 67:1-2 (emphasis added).

ii. Meta abandoned licensing efforts not because licensing is untenable, but because of unlawful substitution.

Meta never actually executed any large-scale licenses with those publishers. Meta claims this is because it learned publishers did not have rights for AI licensing, rendering large-scale book licenses infeasible. But as shown extensively in Plaintiffs' Motion, Meta decided to employ the alternative strategy of using pirated versions of books simply to avoid paying for books. Ex. 52 at -135 ("if we license once [sic] single book, we won't be able to lean into the fair use strategy."). This is a textbook example of the direct market substitution that "underscores the non-transformative nature of [Meta's] use" and creates a dangerous likelihood "that cognizable market harm to the originals will occur." *McGucken*, 42 F.4th at 1164 (quoting *Monge*, 688 F.3d at 1182-83) (quotations omitted).

Meta has now resumed some efforts to license books for LLM training. But just like in 2023, Meta is still using its existing repositories of pirated books to decide which additional books to pirate and which to license. Ex. 97 at -196-97 (describing "gap approach" of only licensing content that is not already contained in Anna's Archive). Using their pirated datasets as a guide, Meta is targeting ***only*** the publishers that possess books Meta has not been able to pirate. *Id.* at -198 ("**will only** go after the publishing companies with the largest delta of missing content" between their offerings and what Meta "collected from Anna's Archive & spidermate") (emphasis

[15] Ex. 134, [redacted] Ex. 135, [redacted]

in original); Ex. 75 (spreadsheet of titles from [redacted] and link for copy on LibGen with count of "match," "miss" and "share" of works compared to LibGen). And just as it did in 2023, Meta continues to pirate as soon as it encounters any licensing hurdles. *E.g.*, Ex. 136, Meta_Kadrey_00237702 (April 2024: "it is unclear whether [Internet Archive] would share the underlying in-copyright books collection even under a deal. So you should grab from Anna's [Archive] and not worry about overlap.").[16]

iii. Meta's own documents show it understands licensing to be feasible.

A single document—Meta's Llama 4 licensing strategy deck—completely destroys Meta's position on the non-existence of an AI training data market. In that June 2024 document, Meta adopted a licensing strategy aimed at acquiring "around ~10-20% of total text data corpus from Llama-4 [] ***from licensing***" and noted that further book acquisition "will be heavily reliant on licensing for success." Ex. 97, at -192, -194 (emphasis added). That deck also shows Meta extensively evaluating the market for AI training data licensing opportunities. *Id.* at -197 ("Market analysis reveals that our competitors are securing large partnerships and making significant investments in data acquisition for LLM training.") Meta extensively surveyed the training data licensing agreements that its competing LLM developers already executed, including with text publishers such as *News Corp*, *Financial Times*, and *The Atlantic*. *Id.* at -200-01. Meta then listed numerous literary publishers that it could license from and estimated the price of those licenses. *Id.* at -202-03. Ironically, one of Meta's "Purchase Criteria" is to "Ensure company has legal right and proper consents (from publishers or authors) to license or sub-license data to Meta." *Id.* at -204. Sophisticated companies like Meta do not conduct market analyses of nonexistent markets.

There is more corroborating evidence. Meta estimates that OpenAI alone is currently spending well over ***$230 million annually*** on licensing data from collective copyright holders. Ex. 138, Meta_Kadrey_00172888, at -905. Moreover, just seven months before Meta filed its Motion

[16] IA contains a mix of copyrighted and public domain material. Its copyrighted contents are also included on Anna's Archive, which Meta downloaded after discussions with IA moved too slowly for Meta's liking.

arguing to this Court that any "theoretical market for licensing text as training data is doomed," its employees circulated a licensing strategy update proclaiming that "the market for licensing for the AI training use case is emerging." *Id.* That document also details the progress of active licensing negotiations for text data with 14 collective rights holders, including books publishers [redacted]. *Id.* at -893-99. Meta estimated that the "typical timeline to closing a licensing deal" is approximately six weeks—nowhere near an impenetrable barrier to data acquisition. *Id.* at -901-02 (reflecting a "Typical Deal Schedule" table). Meta's understanding of the feasibility of the market is also corroborated by external evidence. [redacted]—another LLM competitor—recently signed a deal with [redacted] to collectively license books as LLM training data. Ex. 139, [redacted][17] And in the few weeks since Meta filed its brief, yet another collective book licensing arrangement for AI training was publicly reported.[18]

iv. The transaction costs of collective licensing are not prohibitive.

In the face of this overwhelming evidence of an *actual* market for licensing books as AI training data, Meta is left to repeat an *ipse dixit* argument that negotiating the rights with publishers and authors is too hard. Meta ignores that book publishers—even when they do not hold the collective rights—"can act as intermediaries in the market for LLM training data and can obtain consent from their authors and negotiate royalties for collective licenses." Ex. 126 ¶ 174.[19] Meta's argument is really just an observation that *some* transaction costs would accompany licensing efforts. This reflects the unremarkable fact that that publishers would need to first coordinate with authors for the assignment of rights. Ex. 141, Meta_Kadrey_00152927 ([redacted] explaining

[17] Meta's argument that the [redacted] is irrelevant is yet another factual dispute. Its own expert acknowledged that identical comparators are not typical in IP licensing. Ex. 140, Bakewell Tr. at 264:4-14.

[18] Sam Quigley, *News/Media Alliance Announces AI Licensing Partnership with ProRata*, EDITOR & PUBLISHER (Mar. 27, 2025), https://www.editorandpublisher.com/stories/newsmedia-alliance-announces-ai-licensing-partnership-with-prorata,254978.

[19] For this reason, Meta's argument that the licensing market is thwarted by low "marginal value" of individual books also fails. Br. at 28. Again, it ignores the market reality of collective licenses. *Cf. BMG Music v. Gonzalez*, 430 F.3d 888, 891 (7th Cir. 2005) ("copiers such as [defendant] cannot ask courts (and juries) to second-guess the market and call wholesale copying 'fair use'").

to Meta that they can license some books independently and coordinate with authors for others). The insinuation that the mere existence of transaction costs renders an entire market infeasible is certainly false—every market has a mix of willing buyers and sellers at varied price points. Meta also presupposes it has some absolute right to train Llama on any content it desires, irrespective of the appetite of the content owner. Meta has no such right. Instead, it has the right to participate in the established market for LLM training data (and book markets more generally), where it can make business decisions on what content to license and how much to pay to acquire those licenses from willing licensors.

History also shows that once online piracy is judicially prohibited, market forces naturally react by developing legitimate alternatives. Take Napster, for instance. Shortly after Napster was enjoined, "record companies [developed] license agreements or joint ventures with other Internet companies to distribute their music." 1 *Lindey on Entertainment, Publ. & the Arts* § 2:28 n. 36 (3d ed. 2024). Apple's iTunes proliferated immediately in Napster's aftermath, "allow[ing] users to legitimately download music at 99 cents per song." *Id.* The lesson is clear: once major participants in pirated markets are *forced* to use legitimate alternatives to obtain copyrighted content, those markets subsequently develop, even where collective licensing is necessary. In light of the already-growing market for licensing books as AI training data, there is little reason to believe a similar result is *impossible* here—which is what Meta claims to be the case, and what Meta needs to prevail as a matter of law.

v. Meta's cases are easily distinguished.

Even if Meta were somehow correct that this massive licensing market does not exist or is categorically unavailable to Plaintiffs, Meta's cases are readily distinguishable because they involved circumstances lacking any evidence of even a potential licensing market for the infringed works for the uses at issue. *Tresóna Multimedia, LLC v. Burbank High Sch. Vocal Music Ass'n* stands for the unsurprising proposition that the use of 20 seconds of a copyrighted song in the context of "nonprofit show choir performances" would not displace the market for the entire song. 953 F.3d 638, 652 (9th Cir. 2020). *Seltzer v. Green Day, Inc.* is also of no help to Meta. 725 F.3d

1170 (9th Cir. 2013). There, the court could not determine harm to a licensing market because the plaintiff "provide[d] no additional information" to demonstrate a developing market. *Id*. at 1179.[20] Finally, Meta's heavy reliance on *Bill Graham Archives* is misplaced because in that case, there was no evidence that the copyright holder "lost license revenue from the *uses at issue here*." 448 F.3d at 615 n.6. (emphasis added). That is decidedly not the case here, where Meta pirated exact copies of Plaintiffs' Books and used them as AI training data rather than licensing the Books as AI training data.

b. Meta's conduct harms Plaintiffs' book sale market.

Meta concedes there is an established market for purchasing Plaintiffs' copyrighted books. *See e.g.,* Ex. 142, Sinkinson Report ¶ 51. At minimum, Meta's unauthorized copying deprived Plaintiffs of sales revenue for copies of their Books that Meta made without authorization. *See* Ex. 3; Dkt. 485, ¶ 100 (Meta admits it did not seek or obtain Plaintiffs' permission). There is concrete evidence of actual harm from both Meta's specific piracy and widespread piracy by others: it costs the publishing industry hundreds of millions of dollars per year.[21] The harms attendant to stealing copyrighted works instead of paying for them are obvious and intuitive. *See*, *e.g.*, *Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith*, 11 F.4th 26, 50 (2d Cir. 2021) ("That harm [the destruction of the broader market if the copying Warhol engaged in were to become widespread] is . . . self-evident.") *aff'd* 598 U.S. 508 (2023); *McGucken,* 42 F.4th at 1163 (explaining that substitution "if carried out in a widespread and unrestricted fashion . . . would destroy [plaintiffs'] licensing market"); *Hachette*, 115 F.4th at 193 (noting that courts do not require empirical data to

[20] Meta claims—contrary to precedent—that Plaintiffs not having licensed their materials in the first instance precludes them from asserting a loss in the developing market, even though their cited cases do not support such a conclusion. *See Video-Cinema Films, Inc. v. Cable News Network, Inc.*, 2001 WL 1518264, at *8 (S.D.N.Y. Nov. 28, 2001) (comparing its use of entire works to the use of clips that were "too few, too short, and too small in relation to the whole," in a case without evidence of a market); *Kienitz v. Sconnie Nation LLC*, 766 F.3d 756, 758 (7th Cir. 2014) (plaintiff would have been able to license photograph for apparel whereas Meta's conduct obliterates licensing opportunity if practice widespread among LLM developers).

[21] *See* DAVIS, CHERYL L. & KAZI, UMAIR, PIRACY OF BOOKS IN THE DIGITAL AGE, IN THE ROUTLEDGE COMPANION TO COPYRIGHT AND CREATIVITY IN THE 21ST CENTURY, 21 (BOGRE, MICHELLE AND WOLFF, NANCY EDS., 2020) (noting that in 2017, the publishing industry lost over $315 million dollars due to pirated direct substitution).

support the "logical inference[]" that approving of infringing use leaves "little reason for consumers . . . to pay . . . for content they could access for free").[22]

Meta's arguments thus ignore both reality and the evidence in this case: Meta's theft unquestionably cost each Plaintiff *at least* one book sale—the lost sale *to Meta*. Ex. 126 ¶ 273 ("Meta deprived authors of revenues from book sales by making unauthorized copies of their works."). Straightforward substitution inflicts an obvious harm to the purchasing and licensing markets for Plaintiffs' works. *See Campbell*, 510 U.S. at 591 ("when a commercial use amounts to mere duplication of the entirety of an original, it clearly supersedes the objects . . . of the original and serves as a market replacement for it, making it likely that cognizable market harm to the original will occur") (cleaned up).

Meta offers no evidence to refute the lost sales resulting from its pirated acquisition of Plaintiffs' works. Instead, Meta focuses only on whether users deployed Llama to write books that competed with Plaintiffs'. Its assertion that Plaintiffs admitted they did not lose sales or licensing revenue is misleading. Meta's RFAs asked Plaintiffs to admit they were unaware of lost sales "*other than* [their] contention[s] that LLM developers such as Meta should have compensated [them for the] alleged use [of their] Asserted Works to train large language models." Ex. BG 8 (emphasis added). Meta qualified these RFAs because Plaintiffs uniformly denied earlier unqualified RFAs regarding lost sales—stating that their works were "copied by Meta without consent, without credit, and without compensation." Ex. 4. The relevant question is about lost sales *related to Meta's piracy* and use of their books for training without consent, which several Plaintiffs pointed out in their depositions.[23] Notably, Meta's experts ignored lost sales from Meta's

[22] Though Plaintiffs' infringement claims are not based on Llama's outputs, Dr. Spulber notes that Llama's output harm authors, such as "depriving Plaintiffs of revenues by using copyrighted books as inputs to allow the creation of works that could compete with Plaintiffs' works." Ex. 126 ¶ 192. Dr. Spulber points to the "AI-generated copycat books, often of low quality, flooding Amazon and affecting demand for authors' books." *Id*. ¶ 199.

[23] *See*, *e.g.*, Ex. 145, Klam Tr. 335:9-17 ("Q. Are you aware of any instance in which someone chose not to take a license from you for AI training because of something Meta did? A. I'm not sure where we are here. ***No one has offered to license any of my work for AI training. Q. Okay.***

piracy. Ex. 144, Sinkinson Tr. 257:2-11 (admitting he did not consider "the impact of piracy on book sales"); Ex. 140, 291:11-18 (Bakewell relying entirely on Sinkinson's book sales analysis).

Meta's focus on whether Llama users create competing works is misplaced. But there are disputed facts about that too. Dr. Spulber observed the influx of "AI-generated copycat books, often low quality, flooding Amazon and affecting demand for authors' books." Ex. 126 ¶ 199, *see also* ¶¶ 200-203. Meta's expert claimed to offer opinions about Plaintiffs' book sales in the eight weeks after the Llama 3's release, but he only reviewed book sale ranks, not actual book sales. Ex. 142 ¶¶ 93, 283. He later admitted the impact of Llama 3 would not even be fully observable within the eight weeks he considered. Ex. 144, 76:16-21. Such a methodologically flawed analysis of book *ranks* on a single website over the span of just eight weeks cannot offer any meaningful conclusions about Plaintiffs' book sales, much less wrest a critical fact from the jury.

c. Public benefit considerations support Plaintiffs.

Meta sidesteps the effect widespread piracy would have on the licensing market and focuses instead on the benefit from AI. Meta flatly misstates the proper analysis. When weighing public benefit, courts are to consider the benefit achieved from the copying itself. *Oracle*, 593 U.S. at 35 ("[W]e must take into account the public benefits *the copying* will likely produce.") (emphasis added); *see also New York Times Co. v. Microsoft Corp.*, 2024 WL 4874436, at *3 n.3 (S.D.N.Y. Nov. 22, 2024) ("[A] discussion of 'public benefits' must relate to the benefits from the copying."). There is no public benefit to Internet piracy, and Meta's own employees had serious misgivings over using pirated works. Ex. 32 ("I feel that using pirated material should be beyond our ethical threshold.").

Meta's employees were right to identify the dissonance between Meta's claims and its actual practices. In analyzing the Fourth Factor, courts consider whether the conduct of the alleged infringer—if unrestricted—would "create incentives to pirate intellectual property." *Monge*, 688

A. That's why we are suing you guys.") (emphasis added), Silverman Tr. 204:20-23 ("Q. Are you aware of any instance in which somebody wanted to pay to license use of your work but chose not to because of something Meta did? A. ***Meta itself*** and ChatGPT.") (emphasis added), Kadrey Tr. 222:10-15 (Meta's attorney acknowledging lost licensing opportunity to Meta).

F.3d at 1182; *Dr. Seuss*, 983 F.3d at 461 (finding defendant's unrestricted and widespread unauthorized use of *Oh, the Places You'll Go!* in derivative work "could create incentives to pirate intellectual property and disincentivize the creation of illustrated books") (quotation marks omitted). It is undisputed that Meta avoided paying any of the Plaintiffs a purchase price or a licensing fee by pirating their works instead. And it is self-evident that the market to license Plaintiffs' works as training data would be fatally undermined if all LLM companies followed suit. *See generally Hachette,* 115 F.4th at 163 (noting courts do not require empirical data to support the "logical inference[]" that approving of infringing use leaves "little reason for consumers . . . to pay . . . for content they could access for free"). Plaintiffs cannot "compete with free." *Id*. at 190.

For this reason, approving Meta's conduct would incentivize far-reaching piracy of copyrighted works. Meta torrented hundreds of copies of Plaintiffs' works from multiple established black-market websites. A determination that such conduct is fair use will incentivize all LLM companies—present and future—to pirate. It will also incentivize LLM companies to support and defend such shadow libraries that continue to make stolen works available for free.

Finally, the speculative public benefits of Llama—*e.g.*, providing a platform that enables innovative and "life-saving services and technology"—would also appear to provide a great commercial benefit to Meta. Br. at 32. Meta's desire to find profitable commercial applications for Llama cannot justify its refusal to pay for the works used to help build it. *See generally Harper & Row*, 471 U.S. at 559 ("To propose that fair use be imposed whenever the social value of dissemination outweighs any detriment to the artist, would be to propose depriving copyright owners of their right in the property precisely when they encounter those users who could afford to pay for it.") (quoting Gordon, *Fair Use as Market Failure: A Structural and Economic Analysis of the Betamax Case and its Predecessors*, 82 COLUM. L. REV. 1600, 1615 (1982)) (quotation marks omitted).

III. META'S INTENTIONAL CMI STRIPPING VIOLATED THE DMCA

The record is replete with evidence that Meta's intentional removal of CMI was done "knowing . . . [or] having reasonable grounds to know" that this conduct would "conceal an

infringement." 17 U.S.C. § 1202(b)(1). No more is needed to establish liability under the DMCA. *See Stevens v. Corelogic, Inc.*, 899 F.3d 666, 674 (9th Cir. 2018) (Section 1202(b)(1) requires a showing "that the defendant was aware or had reasonable grounds to be aware of the probable future impact of its actions."). Summary judgment on the DMCA claim should be denied due to the existence of significant material facts in dispute regarding *why* Meta knowingly removed CMI.

A. There Is No Dispute That Meta Intentionally Removed CMI From Copyrighted Books.

Meta does not dispute it intentionally removed "Copyright Management Information" ("CMI")—including the title, date, author's name, and copyright notice—from millions of copyrighted works it copied from pirated databases.[24] Meta Br. at 38 (citing Meta witnesses' testimony about its CMI removal); Answer, Dkt. No. 485, ¶ 89 (admitting Meta ran a script designed to remove from training data "lines that contain the word 'copyright'" and other "repetitive text"). To do so, Meta wrote a script that systematically stripped CMI from the first 25% and last 25% of each copyrighted literary work it downloaded from LibGen, including "[r]ows containing . . . ["ISBN", "Copyright", "©", "All rights reserved", "DOI"]." Ex. 56 at -798-99; *see also* Ex. 146, Krein Opening Report, ¶¶ 146-150. Meta also stripped CMI from *all* of the pirated books in its Books3 dataset. Ex. 147, Lopes Opening Report, ¶¶ 148-151. Meta thus concedes its removal of CMI was intentional.

B. There Is A Genuine Dispute Whether Meta Concealed Copyright Infringement.

Meta argues that CMI is "duplicative, boilerplate text" and that its systematic stripping of such text from copyrighted pirated works was aimed only to "enhanc[e] the overall quality of the datasets," rather than conceal Meta's use of copyrighted works. Dkt. 494, Bashlykov Decl., ¶ 9; Meta Br. at 38-39. Facts in the record contradict that claim.[25]

[24] *See* 17 U.S.C. § 1202(c) (defining "CMI").

[25] Further, to the extent that the Court decides the use of copyrighted material by the LLM constitutes copyright infringement, that serves as a concession that Meta's CMI stripping also "facilitated" such infringement in violation of the DMCA regardless of Meta's motive in doing so.

1. Meta's CMI Removal Concealed the Copyrighted Nature of its Training Data.

Meta's CMI removal was one of several tactics Meta used to conceal from the public (including rightsholders) its use of pirated works in training data due to the legal, regulatory, and reputational risks of training on such data. *See, e.g.*, Ex. 61 at -246 ("If there is media coverage suggesting we have used a dataset we know to be pirated, such as LibGen, this may undermine our negotiating position with regulators"); *see also id.* at -245 ("in **no case** would we disclose publicly that we had trained on libgen") (emphasis in original). Indeed, Meta ceased identifying *all* sources of its training data after its Llama 1 publication, which disclosed Meta's use of Books3. Pltfs' Br. at 7; Ex. 125, 24:21-26:15 (noting datasets used to train Llama 2 and 3 were not disclosed publicly, which was "a decision by our legal department" and LeCun had "no input").

Within months of that release, the pirated origins of Books3 were chronicled in an *Atlantic* exposé, causing Meta employees to discuss whether the public could "deduce" that Meta continued to use Books3 with Llama 2. Ex. 156, Meta_Kadrey_00054905, at -905; Ex. 148, Meta_Kadrey_00063472, at -475 (stating in response to *The Atlantic* exposé on Books3 that Meta's continued use of Books3 was not shared externally). In addition to no longer disclosing its training data, Meta finetuned its Llama models to state they weren't trained on pirated copyrighted data when prompted by end users. Ex. 149, Meta_Kadrey_00054518, at -518-524 (discussing streamlining finetuning strategies to prevent public exposure of Meta's use of pirated data); Ex. 150, Touvron Tr. 403:11-404:7 (discussing Meta AI responding "no" to a query whether it was trained on LibGen and recognizing the answer it was trained to give was false).

CMI stripping was another tool in Meta's arsenal to conceal its use of pirated training data. By removing CMI from works still under copyright, Meta sought to reduce the chance that CMI, such as the © symbol, could be regurgitated, which would alert the public to the use of copyrighted material. Meta also tried to reduce the risk of its models memorizing and regurgitating copyright protected data verbatim. That LLMs memorize their training data is a well-known phenomenon. *See, e.g.*, Ex. 151, Ungar Opening Report, ¶ 199 ("LLMs tend to memorize facts, phrases, and

texts that appear very frequently in their training data"). Meta extensively studied "memorization" tendencies in its models, using adversarial prompts and other approaches to test its models' propensity to regurgitate training data. Ex. 120, 56:22-57:16 (discussing memorization studies at Meta). Llama's memorization was a serious problem that could block or postpone the launch of a model,[26] and Meta was uniquely concerned about Llama regurgitating two kinds of data in particular: copyrighted data (which Meta often referred to as "IP"[27] data) and data containing "PII," or personal identifying information. Ex. 119 ("Memorization in LLMs & MME"); Ex. 121 at -653 (describing "metrics for text memorization" regarding "PII" and "IP/Copyright (books)"); Ex. 152, Meta_Kadrey_00053071, at –126.00003 ("All datasets approved for use, memorization issue (PII and IP addressed)"). With respect to copyrighted data, Meta developed mitigations that aimed to reduce memorization risks of "IP sensitive data"—in other words, copyrighted books. Ex. 120 101:1–102:1.[28]

As Meta's corporate deponent testified, Meta's removal of "copyright information" was motivated also by "regurgitation concern[s]." *See* Ex. 155, Clark Tr. (Mitigation), 58:21-25. Meta knew CMI carried heightened memorization risks since copyright notices are "repeated substring[s] of data" "prone to [being] memoriz[ed] . . . and then generate[d]." Ex. 156, Meta_Kadrey_00054905, at -906; *see also* Ex. 120, 72:3-19 (testifying Meta was "concern[ed]" about "the repetitive nature of . . . [copyright] notices" and removed them "to avoid kind of the model, like, regenerating text that it had seen before"); Ex. 155, 43:24-44:3 ("among the data being removed . . . [t]he copyright information and line spacing and emails, part of that is for preventing

[26] Ex. 153, Meta_Kadrey_00048554 (discussing "model memorization" with "launch blocking" risks); Ex. 154, Meta_Kadrey_00080247 (discussing whether the company considered "any of the uncommitted mitigations as launch-blocking requirements" and referencing "Copyright/Memorization Mitigations").

[27] Meta consistently used "IP" as a shorthand for "copyrighted books." Ex. 121 at -653.

[28] This included "deduplicating" data (or removing data that appeared repeatedly across datasets) and minimizing "epoching" (or training repeatedly on the same datasets). *See, e.g.*, Ex. 157, Nayak 30(b)(1) Tr. 231:14-233:11 (discussing "copyright memorization mitigations"); Ex. 120, 56:13-20 (identifying "deduplication" as a method for "mitigating regurgitation"); Ex. 158, Meta_Kadrey_00049684, at -689 ("increase epoching on Wikipedia" resulted in Llama models "repeating Wikipedia articles verbatim").

regurgitation"). Notably, Meta's script removed CMI and "PII"—*i.e.*, the very data Meta singled out as legally risky if it's regurgitated. Ex. 56 at -798-99 (documenting the removal of PII and copyright data); Ex. 119 (identifying "PII" and "IP/copyright" as a memorization focus).

2. Meta's Inconsistent CMI Stripping Reveals the Pretextual Nature of its Justification, Further Precluding Summary Judgment.

According to Meta, stripping CMI from copyrighted works in its training datasets served a singular purpose: elimination of "noisy" text from the datasets whose inclusion would hurt model performance. For its argument to hold water, Meta needed to strip the CMI from all books it used for training, whether currently copyrighted or in the public domain. CMI is CMI, after all, and its "noisiness" should not depend on the age of the book. But Meta did the exact opposite. While Meta removed CMI from the pirated works in LibGen and Books3, *it did not strip CMI from public domain books* it used from Project Gutenberg, an online library of free ebooks that are out of copyright. In Meta's "B3G" training dataset,[29] which combines Books3 ("B3") and Project Gutenberg ("G"), every text in Books3 opens with the fields "**footer_strip_count**" and "**header_strip_count,**" indicating the CMI that Meta removed from these works. Ex. 147, ¶ 150. But this (or comparable) language appears nowhere in Meta's copy of public domain works from Project Gutenberg, despite the fact that both sets were processed at the same time. Ex. 159, Clark Tr. Vol. III at 294-295 (discussing data processing of B3G). Instead, the Project Gutenberg books contain their original CMI: approximately 12,000 intact copyright notices. Butterick Decl. ¶ 15; Ex. B (summary table). Meta's decision to remove CMI from books it knew to be pirated and still under copyright, while not removing CMI from books no longer under copyright, clearly raises a triable issue of fact about Meta's "culpable scienter."[30]

---

[29] The B3G dataset is housed within a physical hard drive labeled Meta_Kadrey_Data_001. This hard drive, along with four others, were authenticated as Meta's training data. Ex. 103, 16:15-17:10; 136:20-137:3.

[30] Meta's argument that Plaintiffs' DMCA claim fails if there is a finding of fair use lacks merit. The only court to address this issue held fair use does *not* preclude liability under § 1202(b). *Murphy v. Millennium Radio Grp.,* 2015 WL 419884, at *4-5 (D.N.J. Jan. 30, 2015) ("Congress could have easily drafted § 1202(b) to include a strict resulting infringement requirement, yet it did not.").

An internal debate within Meta to reintroduce "source metadata"—or CMI—back into Meta's LLM training data further undercuts Meta's argument. Ex. 155, 46:23-47:1 (stating that the CMI that Bashlykov removed through his script "was a subset of the overall data that was being tested and evaluated as part of the source metadata strategy"). By returning CMI *back* to works in its LLM training data, Meta wondered if Llama could identify the "source" of information in its outputs. *Id.* 46:11-17 (stating Meta's source metadata project aimed to test whether Meta could know where "knowledge came from" when the model answered the question, "was Abraham Lincoln a president"); Ex. 160, Meta_Kadrey_00054416 (discussing benefits of training on CMI in relation to "better controllability/citations/maybe some positive pretraining effect"). But these potential rewards were offset by the risk of public exposure of this pirated data, which Meta employees described once again as "not acceptable from an IP/brand safety leakage perspective." *Id.*; Ex. 155, 53:16-54:15 (describing "the leakage of source metadata" as the regurgitation of "the exact combination of datasets and/or *data Meta trained on for that model*") (emphasis added). As Meta engineer Melanie Kambadur stated: "we have to be basically 100% confident no one can extract this data, and based on previous memorization work we do not think we can be." Ex. 160. To make training on CMI "safer"—that is, reduce its recognized legal and reputational risk—Meta experimented with encryption, converting CMI into hashes in relation to parts of its web crawled data and its "libgen-fiction" set. Ex. 137, Meta_Kadrey_00179968, at -970-71. Yet even training on encryptions of "titles," "author," and "urls" did not quell Meta's concerns about the "legal issues" if "people . . . figure out the hash associated" with different data on which Meta trained. *Id.* at -971 (Comment 5); Ex. 155, 55:18-25. These facts create a genuine dispute whether Meta removed CMI from copyrighted works to conceal infringement.

IV. CONCLUSION

For the foregoing reasons, Plaintiffs respectfully request that the Court grant Plaintiffs' motion for partial summary judgment and deny Meta's cross-motion for partial summary judgment.

Dated: April 7, 2025

By: /s/ *Maxwell V. Pritt*
Maxwell V. Pritt

LIEFF CABRASER HEIMANN & BERNSTEIN, LLP
Elizabeth J. Cabraser (SBN 083151)
Daniel M. Hutchinson (SBN 239458)
Reilly T. Stoler (SBN 310761)
275 Battery Street, 29th Floor
San Francisco, CA 94111-3339
(415) 956-1000
ecabraser@lchb.com
dhutchinson@lchb.com
rstoler@lchb.com

Rachel Geman (*pro hac vice*)
250 Hudson Street, 8th Floor
New York, New York 10013-1413
(212) 355-9500
rgeman@lchb.com

Kenneth S. Byrd (*pro hac vice*)
Betsy A. Sugar (*pro hac vice*)
222 2nd Avenue South, Suite 1640
Nashville, TN 37201
(615) 313-9000
kbyrd@lchb.com
bsugar@lchb.com

JOSEPH SAVERI LAW FIRM LLP
Joseph R. Saveri (SBN 130064)
Cadio Zirpoli (SBN 179108)
Christopher K.L. Young (SBN 318371)
Holden Benon (SBN 325847)
Aaron Cera (SBN 351163)
601 California Street, Suite 1505
San Francisco, California 94108
(415) 500-6800
jsaveri@saverilawfirm.com
czirpoli@saverilawfirm.com
cyoung@saverilawfirm.com
hbenon@saverilawfirm.com
acera@saverilawfirm.com

Matthew Butterick (SBN 250953)
1920 Hillhurst Avenue, #406
Los Angeles, CA 90027
(323) 968-2632
mb@buttericklaw.com

BOIES SCHILLER FLEXNER LLP
David Boies (*pro hac vice*)
333 Main Street
Armonk, NY 10504
(914) 749-8200
dboies@bsfllp.com

Maxwell V. Pritt (SBN 253155)
Margaux Poueymirou (SBN 356000)
Joshua M. Stein (SBN 298856)
44 Montgomery Street, 41st Floor
San Francisco, CA 94104
(415) 293-6800
mpritt@bsfllp.com
mpoueymirou@bsfllp.com
jstein@bsfllp.com

Jesse Panuccio (*pro hac vice*)
Jay Schuffenhauer (*pro hac vice*)
1401 New York Ave, NW
Washington, DC 20005
(202) 237-2727
jpanuccio@bsfllp.com
jschuffenhauer@bsfllp.com

Joshua I. Schiller (SBN 330653)
David L. Simons (*pro hac vice*)
55 Hudson Yards, 20th Floor
New York, NY 10001
(914) 749-8200
jischiller@bsfllp.com
dsimons@bsfllp.com

Interim Lead Counsel for Individual and Representative Plaintiffs and the Proposed Class

CAFFERTY CLOBES MERIWETHER & SPRENGEL LLP
Bryan L. Clobes (*pro hac vice*)
Alexander J. Sweatman (*pro hac vice*)
Mohammed A. Rathur (*pro hac vice*)
135 S. LaSalle Street, Suite 3210
Chicago, IL 60603
(312) 782-4880
bclobes@caffertyclobes.com
asweatman@caffertyclobes.com
mrathur@caffertyclobes.com

DICELLO LEVITT LLP
Amy Keller (*pro hac vice*)
Nada Djordjevic (*pro hac vice*)
James Ulwick (*pro hac vice*)
10 North Dearborn Street, Sixth Floor
Chicago, Illinois 60602
(312) 214-7900
akeller@dicellolevitt.com
ndjordjevic@dicellolevitt.com
julwick@dicellolevitt.com

David Straite (*pro hac vice*)
485 Lexington Avenue, Suite 1001
New York, New York 10017
(646) 933-1000
dsraite@dicellolevitt.com

COWAN DEBAETS ABRAMS & SHEPPARD LLP
Scott J. Sholder (*pro hac vice*)
CeCe M. Cole
60 Broad Street, 30th Floor
New York, NY 10004
(212) 974-7474
ssholder@cdas.com
ccole@cdas.com

Counsel for Individual and Representative Plaintiffs and the Proposed Class